MECEE 4520 Data Science for Mechanical Systems

Term Project: Clothing Recommendation using Machine Learning

Yingde Xie(yx2552)

Professor. J. Browne

Dec 20th, 2020

Introduction:

With economy and production capability improving for a couple centuries, more and more people are not satisfied about products with limited options, colors, designs to purchase. For instance, in the field of clothing, there are many brands and many designs for only one type of clothing to choose. Since there were so many options for only one type of clothing, the standard of size selection was urgent to be set so people can buy cloths that fits them. Therefore, a set of standard size measurements such as small, medium, large, extra-large in clothing and number 4 to 16 in shoes were invented to assist customer purchasing products.

The history of size system started at the beginning of the 19th century. "The Napoleonic Wars (1803–1815), the Crimean War (1853–1856) and the American Civil War (1861–1865) required unprecedented numbers of uniforms and so full-body sizing systems were developed, making it possible to calculate other parameters based on a single chest measurement." (Team, 2019) As stated above, this was the time when the size of cloths was only based on the measurement of chest. And with our society developing, the size system developed as well. Today, the size system is a lot more complex than before: the length of sleeve, the length of collar, the length of waist area determines the size code of the product, and the physical dimensions of the customers who intend to wear the product determine which code to select. However, selecting the right size for a product is not as simple as it sounds, because products from different companies or designers or usages in the market could have very different size measurements. or instance, I need to buy shoes with size 9 at Adidas and I might need 9.5 for shoes from Nike. Moreover, even in the same brand, products from different designers have different scaling as well. And nowadays, online shopping is the main trend for buying clothing which is impossible to find the best fitting size only by information provided from website.

Therefore, when customer try to purchase these clothing products, there will be a gap between expected satisfaction and the actual satisfaction causing by the size scaling difference. Therefore, instead of looking for a new size system which would cost the entire industry to change, using the computing power we have today to help customers decide what size of a certain product to purchase will be a solution.

In this project, a program will be designed to determine what size to take for a product of user's choices with a series of input parameters from the users. With many Computer vision technologies being revolutionized every day, measuring customer's dimensions will be expected to be provided for the program. And the main idea is that instead of giving instructions like the current size standard is doing, creating a machine learning program that could learn for each product which size is the best fitting size for every single customer's dimensions provide by CV would be helpful, and some other parameters. Even though the idea was to create a program that could cover most of the products in the market, it is critical to narrow the product selections to one hoodie and one T-shirt from *Essential* and one pair of hoodie and T-shirt from *Nike* and the user parameters so that the program would not be too enormous to complete.

Data & Methods:

In this program, a Machine Learning Algorithm called Decision Trees for Classification is the method to train the system and eventually accurately predict good size selections for other users beside the training data. "Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two

entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split." (Blog, 2017) Since the output of the program is to select the size of cloths which are XS, S, M, L, XL, XXL ------ a set of discrete decision variables, the program will use the model of classification decision tree. Such a tree is modeled to go through a series of recursive partitioning until the bottom level and result into one final answer which would be the output in this case the recommended size for the product. An example is shown below:

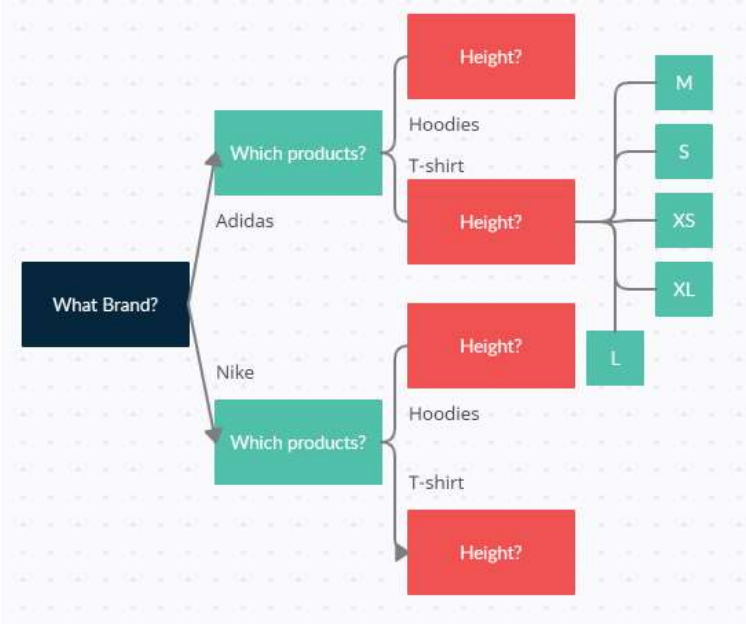

Figure1. example classification decision tree

The Classification decision tree Algorithm for the size selecting problem is following these procedures:

1. Start at the tree root which is selecting the size standard by selecting the exact product from simple yes/no questions.

2. After the standard is assured, split the data on the feature that has the "largest information gain (IG)" (Chakure, 2020) and go to the next level.
3. Iteratively, repeat the splitting process until the feature that has a negligible information gain which means the samples at each leaf nodes are in the same category.

The selected products are: Essentials --- Essentials SS20 Graphic (Hoodie), FW20 Sage(T-shirt) and Nike --- Club Fleece (Hoodie), Sportswear Swoosh(T-shirt). To reduce the complexity of the project, the data about these products are filtered to be in men's session which has a wide range of dimension statistics, so the categorical gaps are more significant. And most of the training data is collected from the comments and sales record of these product on two popular online shopping websites: Dewu and Taobao. And a small part of the training data and the validation data is from personal interview of store owners who have sufficient experience with selecting size for these products. After cleaning the raw data, there are 3,000 samples left for the training

The tool for making the Classification decision tree is a python tool called sklearn tree --- DecisionTreeClassifier which would generate a decision tree based on the features of our selections and result in a determined output which is the size in this case. A major part of the program is determining the features and desired output:

- The selected features are:
    - Brands:
        - Nike
        - Essential
    - Product Types:
        - Hoodies
        - T-Shirt
    - Collar Length needed
    - Chest Length needed
    - Waist Length needed
    - Personal Dressing Style:

        - Prefer Oversize
        - Want perfectly fit
        - Does not care
    - Location of Residence:
        - North
        - South
    - Occupation for the product:
        - for school
        - for partying
        - just casually wear them.
- And the selected output will be the **size selection of that product**:
    - XS, S, M, L, XL, XXL.

Since some of the data are string, which is not acceptable for the sklearn tree DecisionTreeClassifier which is the main tool, one way of fixing this problem is to replace all string variable into numerical numbers, such as Nike is 0, and Essential is 1. Therefore, after the decision tree is generated, with a code hint, we can translate the code back to string content. For instance, on the result, it indicates that as the Occupation is less than 0.5 which is wearing the product for partying and personal style is less than 0.5 which means the customer prefer oversize style, the model would predict the output for size selecting to be one size larger than the one the customer need based on only his dimensions.

After the model is figured, the next step is to export the model into a graph that could better help visualize the relationship between the features and the output. In this program, a tool called graphviz is used. And a png file is saved for better visualization.

Results:

After the Decision Tree model is computed, an image of the tree is created and shown below:

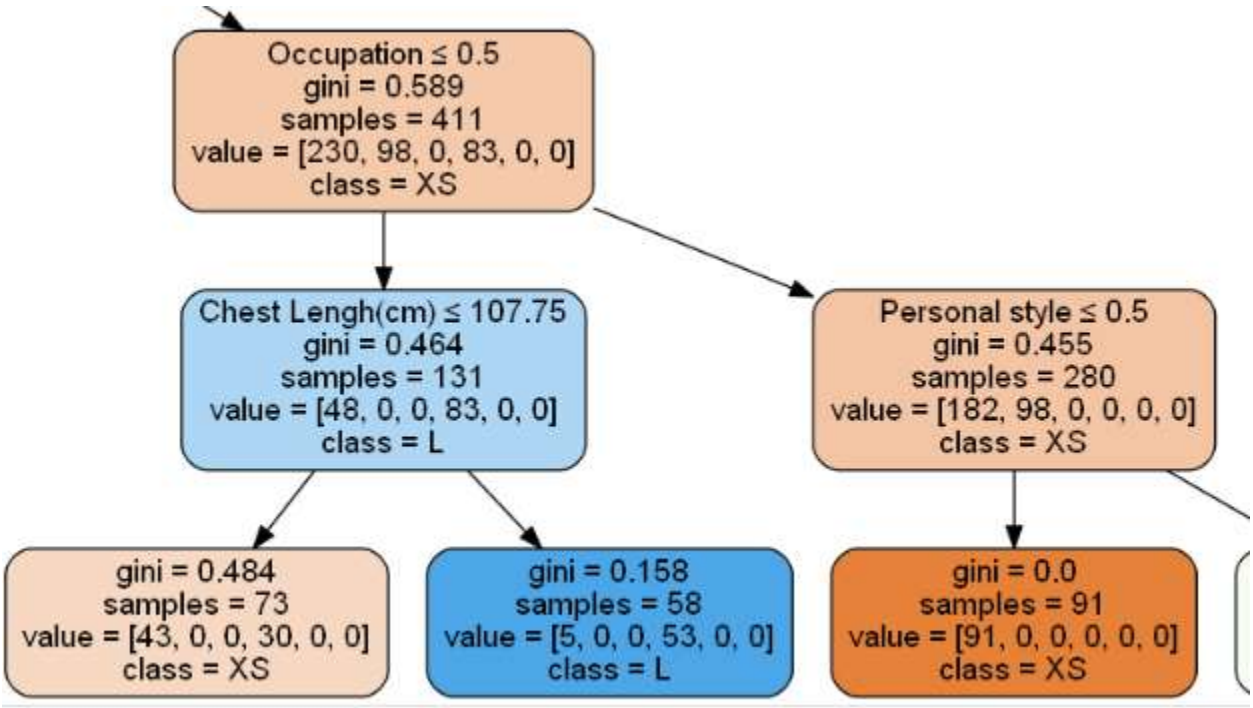

Figure 2. A small section of the final decision tree

After the decision tree model is developed, validations would be needed to see if the program is working correctly. The validations are composed with two parts: first one is an in-sample predictions which are computed using only the training data features and compare the predicted result with training data output; second part is using a complete new set of data with same features and desired output, and then put in the features of this new set of data and compare the final output. The figure below is showing the first part of the validation of the model:

```
First in-sample predictions: ['L' 'M' 'M' 'XXL' 'XXL']
Actual target values for those homes: ['L', 'M', 'M', 'XXL', 'XXL']
```

Figure 3. In sample validation

We can see that the model did a very good job at predicting the desired output from the training data since the predictions match closely.

```
The first five predictions of the model are ['L' 'XS' 'XL' 'L' 'L']
The first five validating outputs are['M', 'XS', 'XL', 'L', 'M']
```

Figure 4. Out sample validation

We can see that the model did not do a good job at predicting the desired output from the new data since the predictions match not quite closely.

```
0.17892857142857144 0.23618090452261306
```

Figure 5. Average error for in-sample validation and out-sample validation

The average error (mistakes divided by number of comparisons) is 0.178 meaning that there are 17.8% chance for the model too recommend a size that is desired for the customer with the training data. And the average error (mistakes divided by number of comparisons) is 0.236 meaning that there are 23.6% chance for the model too recommend a size that is desired for the customer with the new data.

Discussion:

For better visualization, the maximum depth of the example tree is limited to 5 levels. Since there are 8 features controlling the final output, it is critical that the depth of the tree should be at least 8 levels so that each feature would have the opportunity to make influence for the final output. And I believe that with the maximum depth being increased, the accuracy would be increase which as well means that the average error would be small so the system would be able to find an accurate size recommendation for each customer.

For future work, the system could be developed to adapt more features and maybe be developed into an actual product tool for online shopping with CV and some personal survey being the input.

Reference:

1. Santos-Longhurst, A., & Whelan, C. (2019, July 12). Shoe Width: Sizes, Measurement, Foot Issues, and Home Remedies. Retrieved December 19, 2020, from https://www.healthline.com/health/shoe-width
2. Team, S. (2019, November 19). A brief history of sizing systems. Retrieved December 20, 2020, from https://medium.com/sizolution/a-brief-history-of-sizing-systems-aee6bd066834
3. Blog. (2017, September 07). Decision Trees for Classification: A Machine Learning Algorithm. Retrieved December 20, 2020, from https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html
4. Chakure, A. (2020, November 06). Decision Tree Classification. Retrieved December 20, 2020, from https://medium.com/swlh/decision-tree-classification-de64fc4d5aac